Contributed comment on Article by Hahn, Murray, and Carvalho

Daria Bystrova, Julyan Arbel, Thibaud Rahier

Contributed comment on Article by Hahn, Murray, and Carvalho

Daria Bystrova[1] **, Julyan Arbel**[1] **and Thibaud Rahier**[2]

[1] *Univ. Grenoble Alpes, Inria, CNRS, LJK, 38000 Grenoble, France.*
e-mail: [daria.bystrova,julyan.arbel]@inria.fr

[2] *Criteo AI Lab, 38000 Grenoble, France. e-mail:* t.rahier@criteo.com

In this discussion of Hahn et al. (2020), we elaborate on the specific scenario when the selection into treatment depends on the outcome under no treatment. We reproduce Example 1 in the paper with $p = 2$ control variables and $n = 250$ observations under the data generating process

$$Y_i = \mu(x_1, x_2) - \tau Z_i + \epsilon_i, \quad i = 1, \dots, n,$$
$$\epsilon_i \sim \mathrm{N}(0, 1), \quad x_{i1}, x_{i2} \sim \mathrm{Uniform}(0, 1),$$

where Y is a measure of heart distress, Z the treatment indicator, and x_1 and x_2 are two control variables. The treatment effect τ is supposed to be homogeneous and set to $\tau = 1$ here. The prognostic function is set to $\mu(x_1, x_2) := \mathrm{E}(Y \mid x_1, x_2, Z = 0) = \mu(x_1, x_2) = 6\Phi(2(x_1 - x_2)) - 3$.

The authors argue that, under strong confounding, *the estimation of average treatment effect (ATE) using BART can exhibit severe bias*. The BCF model is designed to solve this issue, by including the propensity score $\pi(x)$ (or its estimate) in the set of predictors when learning ATE. In this contributed comment, we have chosen to study the extent to which BCF exhibited smaller bias than BART for different levels of confounding. The confounding 'amount' is controlled by a scalar $\alpha \in [0, 1]$ hyperparameter in our experiments:

$$\pi_\alpha(\mu(x_1, x_2), x_1, x_2) := \alpha \left(0.8\Phi \left(\frac{\mu(x_1, x_2)}{0.1(2 - x_1 - x_2) + 0.25)} \right) + 0.025(x_1 + x_2) \right) + 0.05\beta_\alpha,$$
$$\beta_\alpha := \frac{1}{2}(19 - 17\alpha).$$

For $\alpha = 0$ the propensity score is constant and we are in a randomized controlled trial situation (no confounding); for $\alpha = 1$ the propensity score is roughly equivalent to the one used in Example 1 presented in the paper (high confounding). The rationale behind the choice of β_α is that π_α is required to stay roughly constant for average values of the control variables $x_1 = x_2 = 1/2$; see Figure 1. The $\mathrm{bias}(\hat{\tau}) = \mathrm{E}[\hat{\tau} - \tau]$ is computed as an empirical average over 100 datasets[1].

We observe from Figure 1 that for α close to 0, BART and BCF exhibit similar bias, which is expected since BCF has no regularization-induced confounding

[1] Code is available on Github (https://github.com/dbystrova/bcf_discussion).

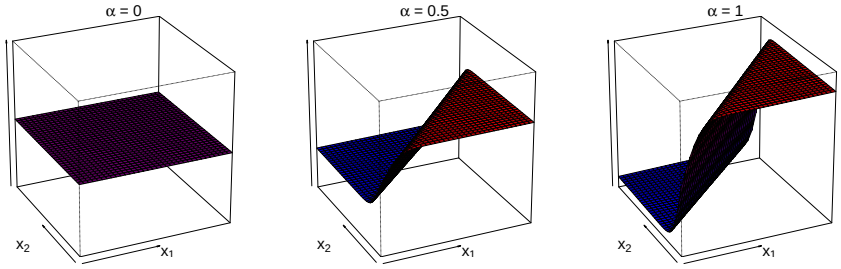

FIG 1. *Propensity function* $\pi(x_1, x_2)$ *for different values of* α *in* $\{0, 0.5, 1\}$.

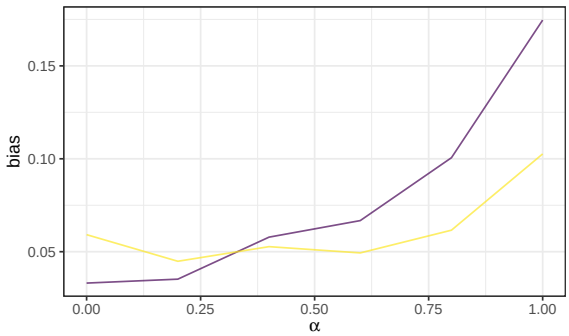

FIG 2. *Bias for BART (purple) and BCF (yellow) models in Example 1 of the paper for varying confounding 'amount'* $\alpha \in [0, 1]$.

(RIC) to correct in that case. Moreover, the more α increases, the larger the bias exhibited by BART compared to BCF. This is consistent with the fact that the bias increase implied by the increase of α is mainly due to the increase in confounding, which BCF was designed to handle better than BART.

The results of our exploratory study are mainly consistent with what was expected upon reading the paper. However, there remains some unanswered questions. 1) Our experiments show that for $\alpha = 0$, BART exhibits slightly lower bias than BCF: this is intriguing as in the case of constant propensity score, both BCF and BART are expected to behave the same in terms of ATE estimation. 2) We notice that the bias exhibited by BCF increases when α increases (even though it does less so than BART's bias). From our understanding, the only additional source of bias the models could suffer from when $\alpha > 0$ compared to $\alpha = 0$ is the result of RIC, which we expect to be fully handled by BCF. These results suggest either that the increase of α induces an other form of bias, or that BCF is not entirely solving the RIC problem.

Note that both of these open questions could be explained by sub-optimal tuning of BART and BCF models in our experiments. We are conscious that our study only scratches the surface of the complicated problem of causal inference and regularization-induced confounding. More experiments are needed to fully

understand the extent to which BCF better handles RIC than BART.

References

Hahn, P. R., Murray, J. S., and Carvalho, C. M. (2020). Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects. *Bayesian Analysis*. 1